CONTACT NAME
Jiri Matela: matela@ics.muni.cz

Real-time JPEG2000 for Digital Cinema and Large-Scale Image Archiving

Jiří Matela[1,2], Martin Jirman[1], Petr Holub[1]

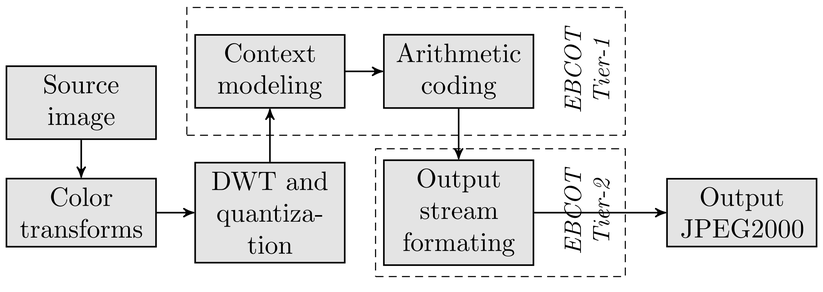
[1] CESNET z.s.p.o., Zikova 4, 162 00 Prague, Czech Republic
[2] Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic

Abstract

JPEG2000 is advanced image compression standard from the JPEG group having applications in various fields, where its processing throughput and/or latency is a crucial attribute and the main limitation of state-of-the-art implementations. JPEG2000 is the compression standard for digital cinema film distribution, it is an emerging standard for broadcast contribution, and it is successfully used for compression of Gigapixel pathology images. Contemporary GPU platforms provide tremendous processing power but they call for highly-parallel algorithm design. JPEG2000 involves several image processing and data compression algorithms, which need specific reformulations in order to efficiently leverage the GPU horsepower. We present successful GPU design and implementation of JPEG2000, allowing for real-time film compression in digital post-production and significant acceleration of compression of large scale pathology imagery.

JPEG2000 Coding Process Overview

The input image data consists of one or more color components which can be optionally transformed into a different color space. Prior to actual compression the data is transformed using Discrete Wavelet Transform (DWT). In case of the lossy compression, the DWT-transformed coefficients are further quantized using a process called uniform scalar dead-zone quantization. Consecutively, the image data is partitioned into code-blocks that are compressed using two-tiered Embedded Block Coding with Optimal Truncation (EBCOT). Each code-block

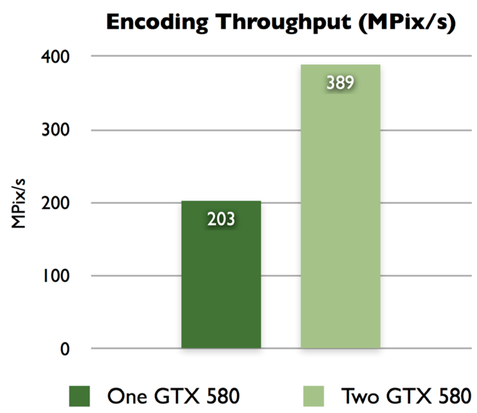


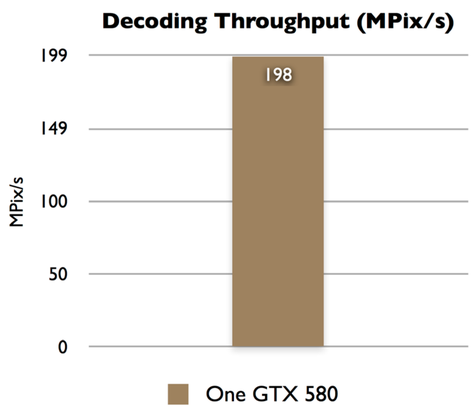
Simplified diagram of JPEG2000 compression work flow

is processed independently by context-modeling and arithmetic MQ-coder in the Tier-1. The context modeler analyzes bit structure of the code-block and collects contextual information, which is then passed together with bit values to the MQ-Coder—a context-adaptive binary arithmetic coder. Compressed bit-stream is finally formated in the Tier-2.

Discrete Wavelet Transform and Quantization

Structure of both DWT and quantization goes along well with the fine-grained data parallelism of the GPU and have been subject to many implementations.

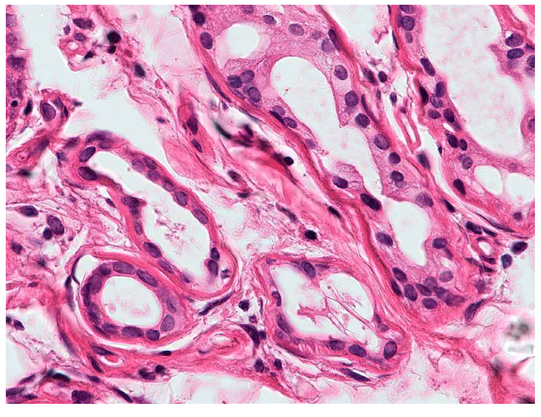




Acceleration of the DWT for GPUs is natural using parallelization of lifting scheme. Image coefficients are updated iteratively using “prediction” and “update” steps, where the number of iterations depends on the width of used DWT filter. The DWT is applied to both rows and columns for 2D image signals. Optimization of memory access is needed in order to avoid problems with non-coalesced access and bank conflicts in shared memory. Quantization can be incorporated as last step of DWT. Our implementation use sliding window approach.

Large-scale Pathology Images

Digital pathology deals with images having resolution of up to several Gigapixels. When compared to traditional image compression formats, JPEG2000 helps to reduce amount of stored data while it maintains image quality. Because of tis region-specific features and direct access to selected parts of the images, it also

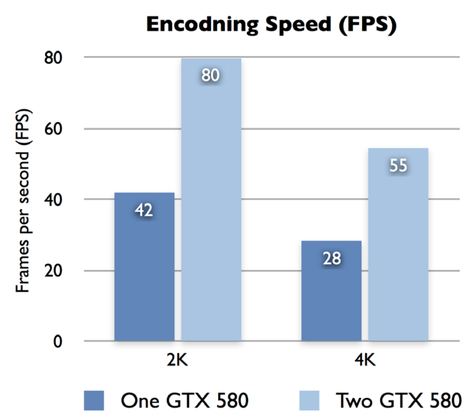


reduces number of files stored, and helps to reduce image data loading latency which improves user experience. The encoding throughput of GPU allows to compress a gigapixel image within about 5 seconds using a single graphics card.

EBCOT Tier-1

EBCOT processes image data by so called code-blocks of 32 × 32 or 64 × 64 pixels, which is rather coarse grained for GPUs. Compression of a code-block exhibits a number of dependencies on the level of individual pixels and even individual bits which limits the parallel processing. We have designed a method that breaks these dependencies and allows for bit-level parallel processing.

- Context modeler codes bits based on three state variables σ, σ’, and η.
- The σ and σ’ indicates that image coefficient is significant.
- The η state is used to determine so called preferred neighborhood (PN) of individual bit positions.
- The σ and σ’ can be precomputed.
  - We compute the PN using a simple fixed-point algorithm.
  - Both steps are done in parallel.
  - Each pixel or bit can be then processed in parallel using the precomputed state information.
  - Arithmetic coder requires careful GPU specific optimizations.



JPEG2000 for Digital Cinema

JPEG2000 is standardized compression for films in digital cinemas in 2K or 4K resolutions. A single NVidia GTX 580 can process a 2K movie at rate of 42 frames per second (fps), which is almost two times the real-time, as the common rate for cinema is 24fps. A 4K movie can be compressed at the rate of 28 fps. The performance scales linearly with number of GPUs.

Results

- GPU can be utilized to accelerate coarse-grained algorithms.
- Algorithm reformulations of Tier-1 is required to utilize parallelism of GPUs.
- GPU specific optimizations are always required.
- Speedups of 7 - 50 times compared to proprietary multi-threaded and single threaded open source CPU implementations.
- Real-time or faster film compression.